

SIMULATION AND ANALYTICAL APPROACH TO THE IDENTIFICATION OF SIGNIFICANT FACTORS

Alexander V. Bulinski¹ and Alexander S. Rakitko¹

Faculty of Mathematics and Mechanics, Lomonosov Moscow State University

Moscow 119991, Russia

bulinski@yandex.ru

Keywords: nonbinary random response; identification of significant factors; regularized estimates of prediction error; exchangeable random variables; central limit theorem.

ABSTRACT

We develop our previous works concerning the identification of the collection of significant factors determining some, in general, non-binary random response variable. Such identification is important, e.g., in biological and medical studies. Our approach is to examine the quality of response variable prediction by functions in (certain part of) the factors. The prediction error estimation requires some cross-validation procedure, certain prediction algorithm and estimation of the penalty function. Using simulated data we demonstrate the efficiency of our method. We prove a new central limit theorem for introduced regularized estimates under some natural conditions for arrays of exchangeable random variables.

1. INTRODUCTION

In a number of models the (random) response variable Y depends on some factors X_1, \dots, X_n . A nontrivial problem is to identify the set of the most “significant factors”. Loosely speaking, for a given $r < n$ one can try to find such collection $\{k_1, \dots, k_r\} \subset \{1, \dots, n\}$ that Y depends “essentially” on X_{k_1}, \dots, X_{k_r} and the impact of other factors can be viewed as negligible. Note that the problem of this type is important in medical and

¹The work is partially supported by RFBR grant 13-01-00612.

biological studies where Y can describe the state of a patient health. For instance, $Y = 1$ or $Y = -1$ may indicate that a person is sick or healthy, respectively. Note also that in pharmacological studies the values -1 or 1 of a response variable can describe efficient or inefficient application of some medicine. Thus it is clear that binary response variables play an important role in various disciplines. At the same time it is obvious that more detailed description of experiments can be desirable. In this regard we refer, e.g., to Bulinski and Rakitko (2014) where non-binary response variables were studied.

There exist various complementary approaches concerning the prediction of response variable and selection of significant combinations of factors. Such analysis in medical and biological studies is included in special research domain called the *genome-wide association studies* (GWAS). The problems and progress in this important domain are considered, e.g., in Moore et al. (2010) and Visscher et al. (2012). Among powerful statistical tools applied in GWAS one can indicate the principle component analysis (Lee et al. (2012)), logistic and logic regression (Schwender and Ruczinski (2010), Sikorska et al. (2013)), LASSO (Tibshirani and Taylor (2012)) and various methods of statistical learning (Hastie et al. (2008)). Note also that there are various modifications of these methods.

We are interested in the “dimensionality reduction” of the whole collection of factors and so employ the term “MDR method”. This term was introduced, for binary response variable, in the paper Ritchie et al. (2001) and goes back to the Michalski algorithm. However, instead of considering contiguity tables (to specify zones of low and high risk) presented in Ritchie et al. (2001) and many subsequent works we choose another way. Namely, to predict (in general non-binary) Y we use some function f in factors X_1, \dots, X_n . The quality of such f is determined by means of the error function $Err(f)$ involving a penalty function ψ . This penalty function allows us to take into account the importance of different values of Y . As the law of Y and $X = (X_1, \dots, X_n)$ is unknown we cannot evaluate $Err(f)$. Thus statistical inference is based on the estimates of error function. Developing Bulinski et al. (2012), Bulinski (2012), Bulinski (2014) we propose (in more general setting) statistics constructed by means of a prediction algorithm for response variable and K -fold cross-validation procedure.

One of the main results of Bulinski and Rakitko (2014) gives the criterion of strong consistency of the mentioned error function estimates when the number of observations tends to infinity. The strong consistency is essential because to identify the “significant collection” of factors we have to compare simultaneously a number of statistics. Moreover, we proposed in Bulinski (2014) and Bulinski and Rakitko (2014) the *regularized versions* of the employed statistics (involving the appropriate estimates of the penalty function) to establish the central limit theorem (CLT).

The paper is organized as follows. Section 2 contains notation and auxiliary results. In Section 3 we discuss the results of simulations to identify (according to our method) the collection of significant factors determining a binary response variable. In Section 4 we prove the new CLT for our estimates (in general for non-binary response Y) using some natural conditions concerning the arrays of exchangeable random variables.

2. NOTATION AND AUXILIARY RESULTS

Further on we suppose that all random variables under consideration are defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Let Y take values in a finite set \mathbb{Y} which we will identify with the set $\{-m, \dots, m\}$ where $m \in \mathbb{N}$. To comprise binary variables we can assume that their values belong to the set $\{-1, 0, 1\}$ and the value 0 is taken with probability 0. Let also X_1, \dots, X_n take values in an arbitrary finite set $\mathbb{X} = \{0, \dots, s\}$. Choose $f : \mathbb{X} \rightarrow \mathbb{Y}$ and a penalty function $\psi : \mathbb{Y} \rightarrow \mathbb{R}_+$. The trivial case $\psi \equiv 0$ is excluded. Introduce the *error function*

$$Err(f) := \mathbf{E}|Y - f(X)|\psi(Y).$$

It is easily seen that one can write $Err(f)$ in the following way

$$Err(f) = \sum_{y, z \in \mathbb{Y}} |y - z| \psi(y) \mathbf{P}(Y = y, f(X) = z) = \sum_{z \in \mathbb{Y}} \sum_{x \in A_z} w^\top(x) q(z)$$

where $q(z)$ is the z -th column of $(2m + 1) \times (2m + 1)$ matrix Q with entries $q_{y,z} = |y - z|$, $y, z \in \mathbb{Y}$ (the entry $q_{-m,-m}$ is located at the left upper corner of Q),

$$w(x) = (\psi(-m)\mathbf{P}(Y = -m, X = x), \dots, \psi(m)\mathbf{P}(Y = m, X = x))^\top$$

and \top stands for transposition. All vectors are considered as column-vectors. According to Bulinski and Rakitko (2014) we can rewrite $Err(f)$ as follows

$$Err(f) = \sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \psi(y) \mathbf{P}(Y = y, |f(X) - y| > i). \quad (1)$$

The law of (X, Y) is unknown, therefore, for each $f : \mathbb{X} \rightarrow \mathbb{Y}$, we can not evaluate $Err(f)$. Thus it is natural that statistical inference concerning the quality of prediction of the response variable Y by means of $f(X)$ is based on the estimates of $Err(f)$.

Let ξ^1, ξ^2, \dots be a sequence of independent identically distributed (i.i.d.) random vectors having the same law as (X, Y) . For $N \in \mathbb{N}$, set $\xi_N = (\xi^1, \dots, \xi^N)$. We will use approximation of $Err(f)$ by means of ξ_N (as $N \rightarrow \infty$) and a *prediction algorithm* (PA). This PA employs a function $f_{PA} = f_{PA}(x, \xi_N)$ defined for $x \in \mathbb{X}$ and ξ_N and taking values in \mathbb{Y} . More exactly, we operate with a *family of functions* $f_{PA}(x, v_p)$ (with values in \mathbb{Y}) defined for $x \in \mathbb{X}$ and $v_p \in (\mathbb{X} \times \mathbb{Y})^p$ where $p \in \mathbb{N}$, $p \leq N$. To simplify the notation we write $f_{PA}(x, v_p)$ instead of $f_{PA}^p(x, v_p)$. For $S \subset \{1, \dots, N\}$ we set $\xi_N(S) = \{\xi^j, j \in S\}$ and $\overline{S} := \{1, \dots, N\} \setminus S$. For $K \in \mathbb{N}$ ($K > 1$), introduce a partition of a set $\{1, \dots, N\}$ by means of subsets

$$S_k(N) = \{(k-1)[N/K] + 1, \dots, k[N/K]\} \mathbb{I}\{k < K\} + N \mathbb{I}\{k = K\}, \quad k = 1, \dots, K,$$

here $[a]$ is the integer part of a number $a \in \mathbb{R}$. Following Bulinski (2012) we can construct an estimate of $Err(f)$ involving ξ_N , prediction algorithm defined by f_{PA} and K -cross-validation (on cross-validation we refer, e.g., to Arlot and Celisse (2010)). Namely, set

$$\widehat{Err}_K(f_{PA}, \xi_N) := \sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \frac{1}{K} \sum_{k=1}^K \sum_{j \in S_k(N)} \frac{\widehat{\psi}(y, \xi_N(S_k(N))) \mathbb{I}\{A_N(y, i, k, j)\}}{\#S_k(N)} \quad (2)$$

where $A_N(y, i, k, j) = \{Y^j = y, |f_{PA}(X^j, \xi_N(\overline{S_k(N)})) - y| > i\}$. Here, for each $k \in \{1, \dots, K\}$, let $\widehat{\psi}(y, \xi_N(S_k(N)))$ be strongly consistent estimates of $\psi(y)$ (as $N \rightarrow \infty$) for all $y \in \mathbb{Y}$, i.e.

$$\widehat{\psi}(y, \xi_N(S_k(N))) \rightarrow \psi(y) \text{ a.s., } y \in \mathbb{Y}, \quad N \rightarrow \infty.$$

In Bulinski and Rakitko (2014) the criterion was established to guarantee the relation

$$\widehat{Err}_K(f_{PA}, \xi_N) \rightarrow Err(f) \text{ a.s., } N \rightarrow \infty.$$

For $r \in \{1, \dots, n\}$ set $\mathbb{X}_r = \{0, 1, \dots, s\}^r$. Then $\mathbb{X} = \mathbb{X}_n$. We write $\alpha = (k_1, \dots, k_r)$, $X_\alpha = (X_{k_1}, \dots, X_{k_r})$ and $x_\alpha = (x_{k_1}, \dots, x_{k_r})$ where $x_i \in \{0, \dots, s\}$, $i = 1, \dots, n$. In many models it is natural to assume that Y depends only on some collection of factors X_α . We say that a vector α (and the corresponding vector X_α) is *significant* if, for $x \in \mathbb{X}$ and $y \in \mathbb{Y}$, one has $P(Y = y|X = x) = P(Y = y|X_\alpha = x_\alpha)$ whenever $P(X = x) > 0$. In Bulinski and Rakitko (2014) (formula (14)), for each $\beta = (m_1, \dots, m_r)$ with $1 \leq m_1 < \dots < m_r \leq n$, the function f^β was introduced and (formula (19)) prediction algorithm $\widehat{f}^\beta(x, \xi_N(W_N))$ was proposed where $x \in \mathbb{X}$ and $\xi_N(W_N) = (\xi_{n_1}, \dots, \xi_{n_u})$, $W_N = \{n_1, \dots, n_u\} \subset \{1, \dots, N\}$. It was proved (Theorem 2 in Bulinski and Rakitko (2014)) that if $\alpha = (k_1, \dots, k_r)$ is significant then, for any $\beta = (m_1, \dots, m_r)$ and each $\nu > 0$, one has $\widehat{Err}_K(\widehat{f}_{PA}^\alpha, \xi_N) \leq \widehat{Err}_K(\widehat{f}_{PA}^\beta, \xi_N) + \nu$ a.s. for all N large enough. Thus it is reasonable to choose among all $\beta = (m_1, \dots, m_r)$ such vector α that $\alpha = \text{argmin}_\beta \{\widehat{Err}_K(\widehat{f}_{PA}^\beta, \xi_N)\}$ or take for further analysis (using permutation tests, see, e.g., Golland et al. (2005)) several vectors giving the estimated prediction error close to the minimal value. Moreover, for specified sequence $\varepsilon = (\varepsilon_N)_{N \in \mathbb{N}}$ of positive numbers, the regularized versions $\widehat{f}_{PA, \varepsilon}^\beta$ of \widehat{f}_{PA}^β were introduced and the CLT was established (Theorem 3 in Bulinski and Rakitko (2014)) for these estimates. Further extension of such CLT is obtained in Section 4 of the present paper.

3. SIMULATION

To illustrate our approach we consider three examples. For each example we simulated i.i.d. random vectors ξ_1, \dots, ξ_N . Then (for each example) we evaluated the estimate $\widehat{Err}_K(\widehat{f}_{PA, \varepsilon}^\beta, \xi_N)$ where $K = 10$, vector β had appropriate dimension, and for regularization of estimates we employed $\varepsilon_N = N^{-1/4}$, $N \in \mathbb{N}$. After that we took all possible collections β of r factors among n and selected 10 of them with lowest values of estimated prediction error $\widehat{Err}_K(\widehat{f}_{PA, \varepsilon}^\beta, \xi_N)$. For saving time of calculations we used $n = 50$ factors. However the results are interesting and instructive. Let the factors X_i , $i = 1, \dots, n$, be i.i.d. random variables taking values $-1, 0, 1$ with probabilities $1/3$ and Y be a binary response variable with values -1 and 1 . We assume also that r (the cardinality of the collection of significant

factors) is equal to 3 in Example 1 and equals 4 in Examples 2 and 3. In Examples 1 and 2 the impact of the “noise” on response variable is described by means of multiplication of Y by the random variable $(-1)^{Z_\gamma}$ where Z_γ is the Bernoulli random variable, namely, $P(Z_\gamma = 1) = \gamma$ and $P(Z_\gamma = 0) = 1 - \gamma$. We consider $\gamma = 0.1$, that is the mean level of noise is 10%. Assume that Z_γ and $X = (X_1, \dots, X_n)$ are independent.

Example 1. Let $r = 3$ and $Y = Y^0 \cdot (-1)^{Z_\gamma}$ where

$$Y^0 = \begin{cases} 1, & X_2 = 1, X_3 \geq 0, \\ 1, & X_2 = -1, X_3 + X_5 \geq 1, \\ -1, & \text{otherwise.} \end{cases}$$

Here X_2, X_3, X_5 are the factors determining Y .

Example 2. Take $r = 4$ and set $Y = Y^0 \cdot (-1)^{Z_\gamma}$ where

$$Y^0 = \begin{cases} 1, & X_2 = 1, \\ 1, & X_3 + X_5 + X_8 \geq 2, \\ -1, & \text{otherwise.} \end{cases}$$

The factors determining Y are X_2, X_3, X_5, X_8 .

In the following example we consider nonlinear constrains.

Example 3. Let $r = 4$. Set

$$Y = \begin{cases} 1, & 3^{X_1+X_2+X_4} \sin(X_3 Z^{\ln(X_3-2X_4+7)}) > 1, \\ -1, & \text{otherwise,} \end{cases}$$

assuming the random variable Z be uniformly distributed on $[0, 1]$. Let Z and X be independent. Here X_1, X_2, X_3, X_4 are the factors determining Y .

Collections of various factors and corresponding values of $\widehat{Err}_K(\widehat{f}_{PA,\varepsilon}^\beta, \xi_N)$ obtained for $N = 500$ are presented in Tables 1, 2 and 3. Namely, EPE_i stands for \widehat{Err}_K found in the framework of Example i where $i = 1, 2, 3$. Columns n_1, n_2, n_3 (and n_1, n_2, n_3, n_4) in the tables

indicate the choice of factors $X_{n_1}, X_{n_2}, X_{n_3}$ (and $X_{n_1}, X_{n_2}, X_{n_3}, X_{n_4}$), respectively. The same information is provided in Tables 4, 5 and 6 where one has $N = 1000$.

It is worth to emphasize that in all considered examples for large ($N = 1000$) and rather modest ($N = 500$) samples our method permits to identify correctly the collections of significant factors (corresponding to the minimum of prediction error estimates). Moreover, these tables show that the estimated prediction error for significant collections of factors has visible advantage w.r.t. other collections.

n_1	n_2	n_3	EPE_1
2	3	5	0.6336
2	3	32	0.8020
2	3	48	0.8100
2	3	28	0.8260
2	3	4	0.8515
2	3	31	0.8527
2	3	22	0.8528
2	3	34	0.8551
2	3	50	0.8649
2	3	23	0.8652

Table 1: $r = 3$, $N=500$

n_1	n_2	n_3	n_4	EPE_2
2	3	5	8	0.3997
2	3	5	24	0.5901
2	3	5	46	0.5911
2	3	5	32	0.5961
2	3	5	31	0.6014
2	3	5	10	0.6059
2	3	5	14	0.6224
2	3	5	42	0.6250
2	3	5	29	0.6251
2	3	5	22	0.6267

Table 2: $r = 4$, $N=500$

n_1	n_2	n_3	n_4	EPE_3
1	2	3	4	0.0939
1	3	20	42	0.2956
2	3	5	29	0.3211
1	3	4	39	0.3228
1	2	3	8	0.3322
1	3	24	42	0.3355
1	2	3	5	0.3395
1	2	3	20	0.3431
1	2	3	40	0.3487
1	2	3	27	0.3558

Table 3: $r = 4$, $N=500$

n_1	n_2	n_3	EPE_1
2	3	5	0.5675
2	3	32	0.7981
2	3	47	0.8096
2	3	34	0.8126
2	3	4	0.8127
2	3	44	0.8334
2	3	48	0.8369
2	3	22	0.8401
2	3	23	0.8441
2	3	31	0.8442

Table 4: $r = 3$, $N=1000$

n_1	n_2	n_3	n_4	EPE_2
2	3	5	8	0.4768
2	3	5	42	0.6936
2	5	8	11	0.6970
2	5	8	26	0.6974
2	3	5	6	0.6981
2	5	8	12	0.7035
2	5	8	50	0.7039
2	3	5	32	0.7045
2	3	5	27	0.7060
2	3	5	46	0.7063

Table 5: $r = 4$, $N=1000$

n_1	n_2	n_3	n_4	EPE_3
1	2	3	4	0.2278
2	3	4	32	0.3355
1	2	3	6	0.4352
1	2	3	46	0.4663
1	3	4	15	0.4694
1	2	3	27	0.4697
1	2	3	50	0.4704
2	3	4	18	0.4812
2	3	4	44	0.4856
1	3	4	40	0.4862

Table 6: $r = 4$, $N=1000$

However, if N is not large enough the proposed stochastic approach can lead to the choice of a collection of factors which is not (the most) significant. For instance, if $N = 500$ then the right identifications of significant factors have occurred in 99%, 97%, 69% of respective simulations for Examples 1, 2 and 3 (averaging is over 100 performance procedures). In Example 3 this frequency of right identification increases till 93% when $N = 1000$.

Figures 1, 2 and 3 demonstrate for each example the character of stabilization of \widehat{Err}_K fluctuations as N grows. This stabilization of estimates can be explained not only by their strong consistency but also on account of their asymptotic normality. In this regard we concentrate further on the new conditions which guarantee the CLT validity for proposed prediction error estimates.

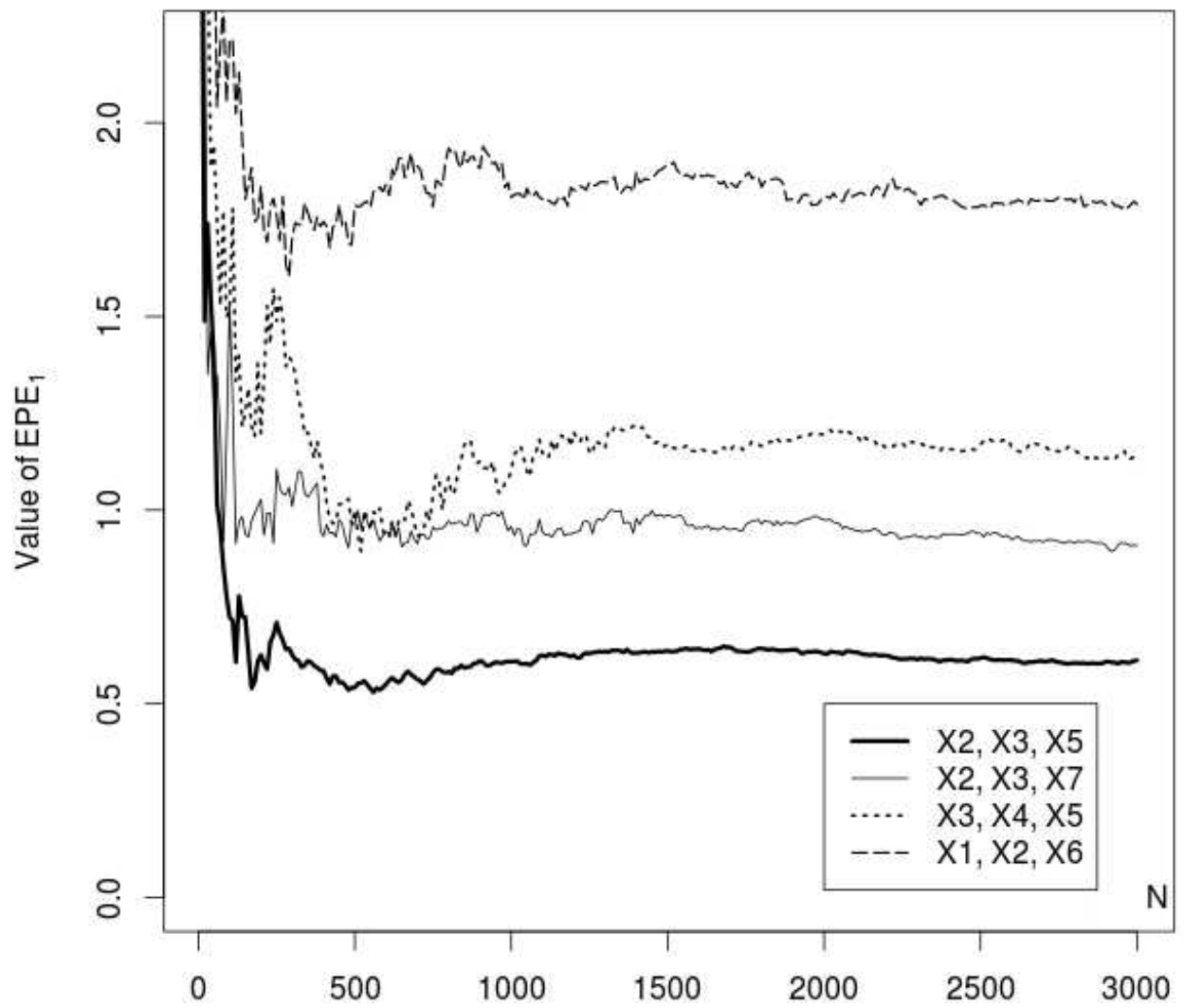


Figure 1: Simulations corresponding to Example 1.

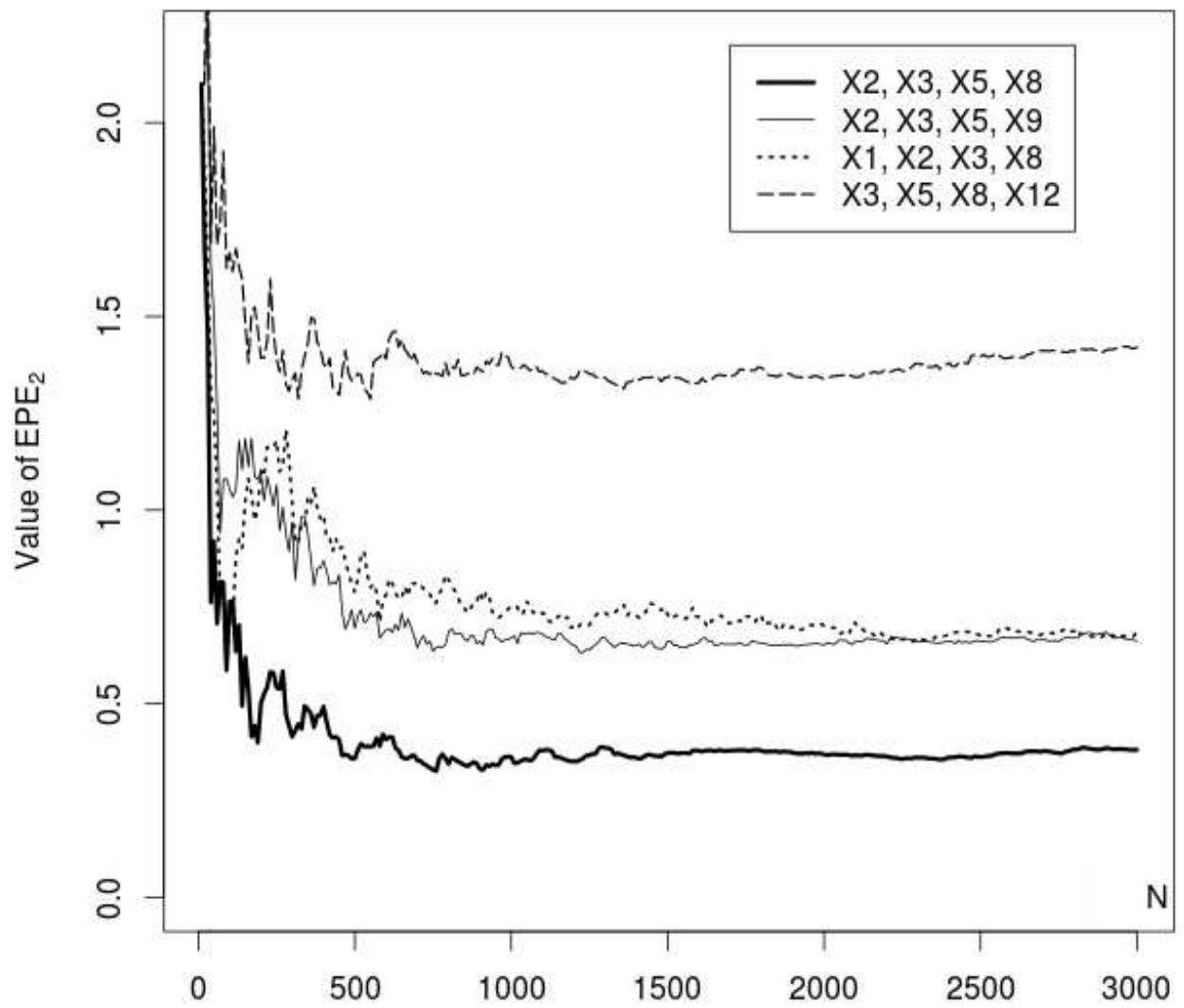


Figure 2: Simulations corresponding to Example 2.

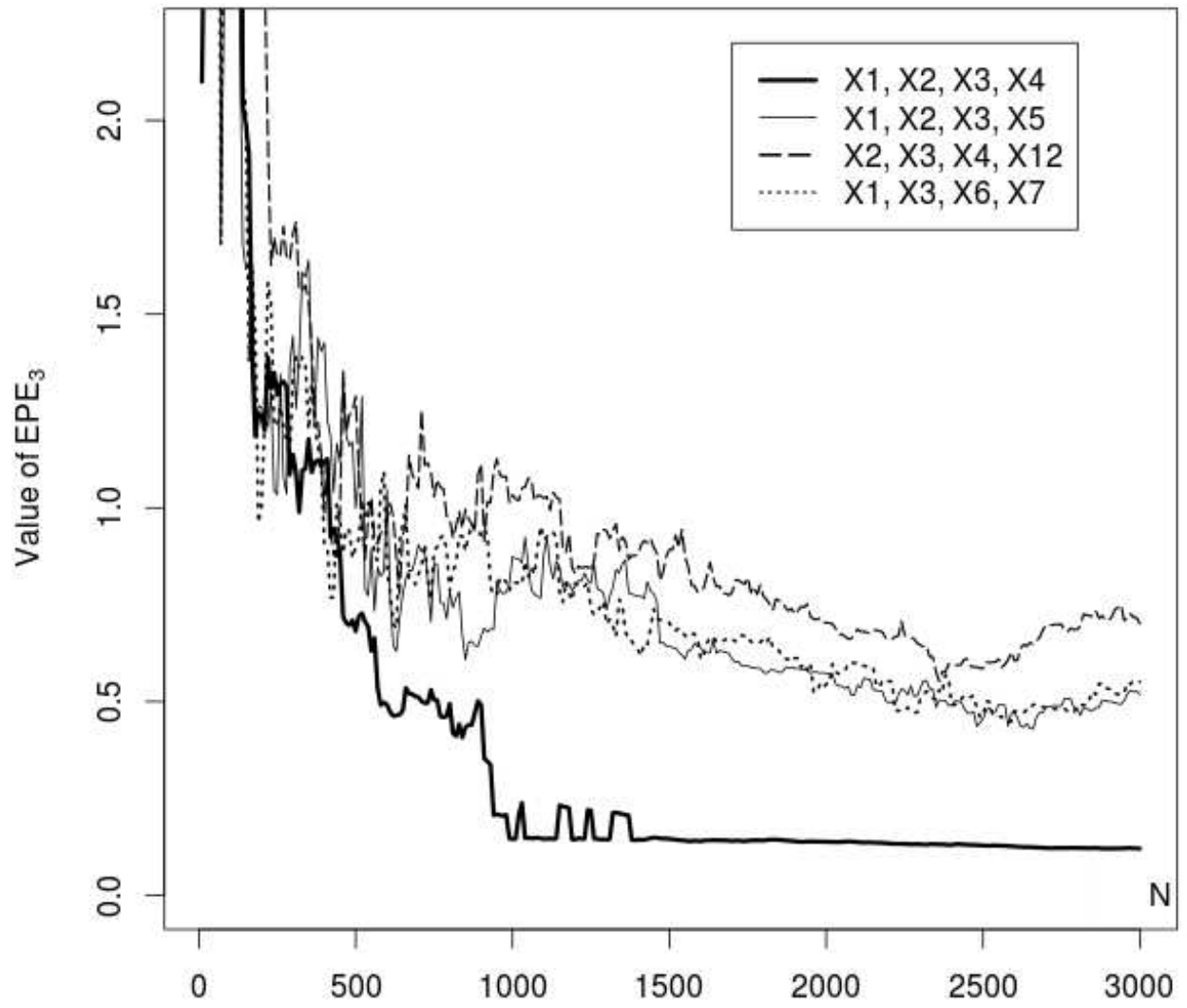


Figure 3: Simulations corresponding to Example 3.

4. NEW VERSION OF THE CENTRAL LIMIT THEOREM

We proved in Bulinski and Rakitko (2014) that asymptotic distribution of random variables $\sqrt{N}(\widehat{Err}_K(f_{PA}, \xi_N) - Err(f))$ coincides with the limit law of

$$\sqrt{N}(\widehat{T}_N(f) - Err(f)) = \frac{\sqrt{N}}{K} \sum_{k=1}^K \sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \frac{1}{\#S_k(N)} \sum_{j \in S_k(N)} h_N(y, i, k, j), \quad (3)$$

as $N \rightarrow \infty$, where

$$h_N(y, i, k, j) = \widehat{\psi}(y, S_k(N)) \mathbb{I}\{Y^j = y, |f(X^j) - y| > i\} - \psi(y) \mathbb{P}(Y = y, |f(X) - y| > i)$$

and $\widehat{\psi}(y, S_k(N)) := \widehat{\psi}(y, \xi_N(S_k(N)))$.

Evidently the summands here are not independent in view of the presence of $\widehat{\psi}(\cdot, S_k(N))$. To prove the CLT for random variables appearing in (3) we used in Bulinski and Rakitko (2014) the hypothesis of asymptotic normality of the vector consisting of two subvectors, one of them being $\sqrt{N}(\widehat{\psi}(\cdot, S_k(N)) - \psi(\cdot))$. Now we employ another approach assuming symmetry of the estimates $\widehat{\psi}(\cdot, S_k(N))$ of a penalty function. Recall the following

Definition 1. A collection of random variables (X_1, \dots, X_n) , $n \in \mathbb{N}$, is called exchangeable if, for any permutation $\sigma \in S(n)$ of the set $\{1, \dots, n\}$, one has

$$Law(X_1, \dots, X_n) = Law(X_{\sigma(1)}, \dots, X_{\sigma(n)}).$$

Take $K \in \mathbb{N}$ and suppose that $N/K = q$ where $q \in \mathbb{N}$. Thus $\#S_k(N) = q$ for each $k = 1, \dots, K$. Consider the sequence of $K \times q$ matrices $(C^{(N)})_{N \in \mathbb{N}}$ with entries

$$\xi_{k,j}^{(N)} := \sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \widehat{\psi}(y, S_k(N)) \cdot \mathbb{I}\{Y^{j+(k-1)q} = y, |f(X^{j+(k-1)q}) - y| > i\} \quad (4)$$

where $k = 1, \dots, K$ and $j = 1, \dots, q$. Introduce

$$X_{N,j} := \frac{1}{\sqrt{K}} \sum_{k=1}^K \xi_{k,j}^{(N)}, \quad j = 1, \dots, q. \quad (5)$$

Then

$$\sqrt{N}(\widehat{T}_N(f) - Err(f)) = \frac{1}{\sqrt{q}} \sum_{j=1}^q (X_{N,j} - \sqrt{K} Err(f)). \quad (6)$$

We take the functions $\widehat{\psi}(y, \cdot)$ which are symmetric for each $y \in \mathbb{Y}$. Then any row and any column of $C^{(N)}$ contain exchangeable random variables (row-column exchangeability). Clearly, the triangular array $\{X_{N,j}, 1 \leq j \leq q, N \in \mathbb{N}\}$ is row-wise exchangeable.

We will establish the CLT for sums appearing in (6). In Berti et al. (2004) one can find several results which guarantee the CLT validity when the summands $\{X_i\}_{i=1}^n$ are (in appropriate manner) conditionally identically distributed. Namely,

$$\frac{1}{\sqrt{n}}(f(X_1) + \dots + f(X_n) - L_n) \xrightarrow{law} Z_{0,\sigma^2} \sim \mathcal{N}(0, \sigma^2) \quad (7)$$

where f is a measurable function such that $\mathbb{E}|f(X_1)| < \infty$ and $L_n = L_n(X_1, \dots, X_n)$. In the mentioned paper the authors applied the martingale techniques. Such approach was developed for exchangeable variables in Weber (1980). We will prove the CLT in the form (7) with $f(x) = x$ for row-wise exchangeable arrays by means of other tools. We will employ the recent result of Röllin (2013). Let $\mathbf{Y} = (Y_1, \dots, Y_m)$ be a collection of exchangeable random variables such that

$$\mathbb{E}Y_1 = 0, \quad \mathbb{E}|Y_1|^3 < \infty. \quad (8)$$

Consider $\Sigma = (\sigma_{i,j})_{1 \leq i,j \leq m}$ with $\sigma_{i,j} = \mathbb{E}(Y_i Y_j)$, i.e. the covariance matrix of \mathbf{Y} . Set $\sigma_{i,i} = \sigma^2$. Suppose that $Y_1 + \dots + Y_m = C_m$ a.s. where C_m is a constant. Then w.l.g. we can assume that

$$\sum_{i=1}^m Y_i = 0 \quad a.s. \quad (9)$$

For a function $h : \mathbb{R}^d \rightarrow \mathbb{R}$ and $k \in \mathbb{N}$ set

$$C_h^{(k)} := \max_{i_1, \dots, i_d \geq 0, \sum_{j=1}^d i_j = k} \left\| \frac{\partial^k h}{\partial x_1^{i_1} \dots \partial x_d^{i_d}} \right\|_{\infty}.$$

Theorem 1 (Röllin (2013)). *Let \mathbf{Y} be a vector consisting of exchangeable random variables and having a covariance matrix Σ . Assume that conditions (8) and (9) are satisfied. Then*

$$|\mathbb{E}h(\mathbf{Y}) - \mathbb{E}h(\mathbf{Z})| \leq C_h^{(2)} \left[\text{var} \left(\sum_{i=1}^m Y_i^2 \right) \right]^{\frac{1}{2}} + 16mC_h^{(3)} \mathbb{E}|Y_1|^3 \quad (10)$$

where $\mathbf{Z} \sim \mathcal{N}(0, \Sigma)$.

For an array $\{X_{n,i}, 1 \leq i \leq k_n, n \in \mathbb{N}\}$ we will use the following notation

$$\hat{\mu}_{k_n} := \frac{1}{k_n} \sum_{i=1}^{k_n} X_{n,i}, \quad \hat{\sigma}_{k_n}^2 := \frac{1}{k_n} \sum_{i=1}^{k_n} (X_{n,i} - \hat{\mu}_{k_n})^2. \quad (11)$$

We apply (10) to prove the following result.

Lemma 1. *Let $\{X_{n,i}, 1 \leq i \leq k_n, n \in \mathbb{N}\}$ be a row-wise exchangeable array where positive integers $k_n \rightarrow \infty$ as $n \rightarrow \infty$. Suppose that*

$$1^\circ. \sup_{n \in \mathbb{N}} \mathbb{E} X_{n,1}^4 < \infty,$$

$$2^\circ. \mathbb{E} X_{n,1}^2 - \mathbb{E} X_{n,1} X_{n,2} \rightarrow \sigma^2 > 0, \quad n \rightarrow \infty,$$

$$3^\circ. \text{cov}(X_{n,1}^2, X_{n,2}^2) + \text{cov}(X_{n,1} X_{n,2}, X_{n,3} X_{n,4}) - 2 \text{cov}(X_{n,1}^2, X_{n,2} X_{n,3}) \rightarrow 0, \quad n \rightarrow \infty.$$

Then, for any sequence $(m_n)_{n \in \mathbb{N}}$ of positive integers such that $m_n \rightarrow \infty$ and $m_n/k_n \rightarrow \alpha < 1$ as $n \rightarrow \infty$, the following relation holds

$$\frac{1}{\sqrt{m_n}} \sum_{i=1}^{m_n} (X_{n,i} - \hat{\mu}_{k_n}) \xrightarrow{\text{law}} Z_{0, (1-\alpha)\sigma^2} \sim \mathcal{N}(0, (1-\alpha)\sigma^2), \quad n \rightarrow \infty.$$

Proof. First of all, for each $n \in \mathbb{N}$, we introduce the auxiliary random variables

$$Y_{n,i} := X_{n,i} - \hat{\mu}_{k_n}, \quad i = 1, \dots, k_n.$$

The collection $\{Y_{n,1}, \dots, Y_{n,k_n}\}$ is exchangeable as $\{X_{n,1}, \dots, X_{n,k_n}\}$ has this property. Obviously $\sum_{i=1}^{k_n} Y_{n,i} = 0$ a.s. for any $n \in \mathbb{N}$. Moreover, $\mathbb{E} Y_{n,1} = 0$, for any $n \in \mathbb{N}$. One can verify that

$$\mathbb{E} Y_{n,1}^2 = \left(1 - \frac{1}{k_n}\right) (\mathbb{E} X_{n,1}^2 - \mathbb{E} X_{n,1} X_{n,2}), \quad \mathbb{E} Y_{n,1} Y_{n,2} = -\frac{1}{k_n} (\mathbb{E} X_{n,1}^2 - \mathbb{E} X_{n,1} X_{n,2}).$$

For each $n \in \mathbb{N}$, take a vector $\mathbf{Z} = (Z_{n,1}, \dots, Z_{n,m_n})$ independent of $(X_{n,1}, \dots, X_{n,k_n})$ and such that $\mathbf{Z} \sim \mathcal{N}(0, \Sigma)$. Here Σ is a covariance matrix of $\mathbf{Y} = (Y_{n,1}, \dots, Y_{n,m_n})$. Thus $\text{cov}(Z_{n,i}, Z_{n,j}) = \text{cov}(Y_{n,i}, Y_{n,j})$, $1 \leq i, j \leq m_n$. Clearly,

$$\frac{1}{\sqrt{m_n}} \sum_{i=1}^{m_n} (X_{n,i} - \hat{\mu}_{k_n}) = \frac{1}{\sqrt{m_n}} \sum_{i=1}^{m_n} Y_{n,i} =: S_{\mathbf{Y}, m_n}.$$

Set $S_{\mathbf{Z}, m_n} := \frac{1}{\sqrt{m_n}} \sum_{i=1}^{m_n} Z_{n,i}$. In view of 2° condition $m_n/k_n \rightarrow \alpha$ ($n \rightarrow \infty$) yields

$$\text{var} S_{\mathbf{Z}, m_n} = \mathbb{E} Y_{n,1}^2 + (m_n - 1) \mathbb{E} Y_{n,1} Y_{n,2} = \left(1 - \frac{m_n}{k_n}\right) (\mathbb{E} X_{n,1}^2 - \mathbb{E} X_{n,1} X_{n,2}) \rightarrow (1 - \alpha) \sigma^2.$$

Consequently, $S_{\mathbf{Z}, m_n} \xrightarrow{\text{law}} \mathcal{N}(0, (1 - \alpha) \sigma^2)$, $n \rightarrow \infty$. Now we show that $S_{\mathbf{Y}, m_n}$ and $S_{\mathbf{Z}, m_n}$ have the same limit distribution. Due to Theorem 7.1 Billingsley (1968) it is sufficient to verify that

$$\mathbb{E} f(S_{\mathbf{Y}, m_n}) - \mathbb{E} f(S_{\mathbf{Z}, m_n}) \rightarrow 0, \quad n \rightarrow \infty, \quad (12)$$

for any three times continuously differentiable function $f : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$c_f^{(j)} := \left\| \frac{d^j f}{dx^j} \right\|_{\infty} < \infty, \quad j = 1, 2, 3.$$

For any fixed $n \in \mathbb{N}$, apply Theorem 1 with $m = m_n$, $Y_i = \frac{1}{\sqrt{m_n}} Y_{n,i}$, $i = 1, \dots, m_n$, and

$$h(x_1, \dots, x_{m_n}) := f(x_1 + \dots + x_{m_n}).$$

Then we can write

$$\begin{aligned} |\mathbb{E} f(S_{\mathbf{Y}, m_n}) - \mathbb{E} f(S_{\mathbf{Z}, m_n})| &= |\mathbb{E} h(\mathbf{Y}) - \mathbb{E} h(\mathbf{Z})| \\ &\leq C_f^{(2)} m_n^{-1} \left[\text{var} \left(\sum_{i=1}^{m_n} Y_{n,i}^2 \right) \right]^{\frac{1}{2}} + 16 C_f^{(3)} m_n^{-1/2} \mathbb{E} |Y_{n,1}|^3. \end{aligned}$$

Note that

$$\begin{aligned} \text{var} \left(\sum_{i=1}^{m_n} Y_{n,i}^2 \right) &= m_n \mathbb{E} Y_{n,1}^4 + m_n(m_n - 1) \mathbb{E} Y_{n,1}^2 Y_{n,2}^2 - m_n^2 (\mathbb{E} Y_{n,1}^2)^2 \\ &= m_n (\mathbb{E} Y_{n,1}^4 - (\mathbb{E} Y_{n,1}^2)^2) + m_n(m_n - 1) \text{cov}(Y_{n,1}^2, Y_{n,2}^2). \end{aligned}$$

We claim that

$$\text{cov}(Y_{n,1}^2, Y_{n,2}^2) - \left[\text{cov}(X_{n,1}^2, X_{n,2}^2) + \text{cov}(X_{n,1} X_{n,2}, X_{n,3} X_{n,4}) - 2 \text{cov}(X_{n,1}^2, X_{n,2} X_{n,3}) \right] \rightarrow 0$$

as $n \rightarrow \infty$. Indeed, set $S_n = \frac{1}{k_n} \sum_{i=1}^{k_n} X_{n,i}$. Using exchangeability property of $(X_{n,1}, \dots, X_{n,k_n})$ and taking into account that covariance function is bilinear we obtain

$$\text{cov}(Y_{n,1}^2, Y_{n,2}^2) = \mathbb{E} Y_{n,1}^2 Y_{n,2}^2 - (\mathbb{E} Y_{n,1}^2)^2$$

$$\begin{aligned}
&= \text{cov}(X_{n,1}^2, X_{n,2}^2) + 2 \text{cov}(X_{n,1}^2, S_n^2) - 4 \text{cov}(X_{n,1}^2, X_{n,2}S_n) \\
&- 4 \text{cov}(X_{n,1}S_n, S_n^2) + 4 \text{cov}(X_{n,1}S_n, X_{n,2}S_n) + \text{cov}(S_n^2, S_n^2).
\end{aligned}$$

For $n \rightarrow \infty$, by virtue of 1° we get

$$\begin{aligned}
\text{cov}(X_{n,1}^2, S_n^2) &= \text{cov}(X_{n,1}^2, X_{n,2}X_{n,3}) + O(k_n^{-1}), \\
\text{cov}(X_{n,1}^2, X_{n,2}S_n) &= \text{cov}(X_{n,1}^2, X_{n,2}X_{n,3}) + O(k_n^{-1}), \\
\text{cov}(X_{n,1}S_n, S_n^2) &= \text{cov}(X_{n,1}X_{n,2}, X_{n,3}X_{n,4}) + O(k_n^{-1}), \\
\text{cov}(X_{n,1}S_n, X_{n,2}S_n) &= \text{cov}(X_{n,1}X_{n,2}, X_{n,3}X_{n,4}) + O(k_n^{-1}).
\end{aligned}$$

Therefore, condition 3° implies that $\text{cov}(Y_{n,1}^2, Y_{n,2}^2) \rightarrow 0$ as $n \rightarrow \infty$. Thus relation (12) holds and the proof is complete. \square

Remark 1. Assume that

$$\sup_{n \in \mathbb{N}} \mathbb{E}((X_{n,1} - \hat{\mu}_{k_n})/\hat{\sigma}_{k_n})^4 < \infty.$$

Then, for a sequence $(m_n)_{n \in \mathbb{N}}$ appearing in Lemma 1, one can prove the following version of the CLT

$$\frac{1}{\sqrt{m_n}} \sum_{i=1}^{m_n} \left(\frac{X_{n,i} - \hat{\mu}_{k_n}}{\hat{\sigma}_{k_n}} \right) \xrightarrow{\text{law}} Z_{0,1-\alpha} \sim \mathcal{N}(0, 1 - \alpha), \quad n \rightarrow \infty.$$

Remark 2. In Chernoff and Teicher (1958) the result similar to Lemma 1 was established but the important case $\alpha = 0$ (which we consider further) was not comprised. One can also employ the martingale approach of Weber (1980) to obtain the result of Lemma 1. However Rollin's Theorem 1 permits us to estimate the convergence rate to the limit Gaussian law. Moreover, we can prove that under certain conditions the asymptotic behavior of the specified partial sums is described by the mixture of the normal laws.

Now we consider the triangular array $\{X_{N,i}, 1 \leq i \leq q, N \in \mathbb{N}\}$ with elements defined by (5). Thus we take $k_n = q$ in Lemma 1 and write N instead of n .

Lemma 2. Suppose that, for each $N \in \mathbb{N}$, any $y \in \mathbb{Y}$ and all $k = 1, \dots, K$,

$$\sup_{y \in \mathbb{Y}, N \in \mathbb{N}, k \in \{1, \dots, K\}} \mathbb{E} \left(\hat{\psi}(y, S_k(N)) \right)^4 < \infty. \quad (13)$$

Let $(m_N)_{N \in \mathbb{N}}$ be a sequence of positive integers such that $m_N \leq q$, $m_N \rightarrow \infty$ and $m_N/N \rightarrow \alpha < 1$ as $N \rightarrow \infty$. Then

$$\frac{1}{\sqrt{m_N}} \sum_{i=1}^{m_N} (X_{N,i} - \hat{\mu}_N) \xrightarrow{\text{law}} Z_{0, (1-\alpha)\sigma^2} \sim \mathcal{N}(0, (1-\alpha)\sigma^2)$$

where μ_N is introduced in (11) (with $k_n = q$ and n replaced by N) and

$$\sigma^2 = \mathbb{E} \left[\sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \psi(y) (\mathbb{I}\{Y=y, |f(X)-y| > i\} - \mathbb{P}(Y=y, |f(X)-y| > i)) \right]^2. \quad (14)$$

Proof. We show that conditions of Lemma 1 are met. 1° follows by virtue of (3), (5) and (13) as indicator function takes values in the set $\{0, 1\}$. Now we turn to 2°. The exchangeability of the columns of the array $\{\xi_{k,j}^{(N)}\}$ implies that

$$\mathbb{E} X_{N,1} X_{N,2} = \frac{1}{K} \mathbb{E} \left(\sum_{k=1}^K \xi_{k,1}^{(N)} \right) \left(\sum_{k=1}^K \xi_{k,2}^{(N)} \right) = \mathbb{E} \xi_{1,1}^{(N)} \xi_{1,2}^{(N)} + (K-1) \mathbb{E} \xi_{1,1}^{(N)} \xi_{2,2}^{(N)}.$$

The Lebesgue theorem on majorized convergence yields that the limit behavior of $\mathbb{E} \xi_{1,1}^{(N)} \xi_{1,2}^{(N)}$ as $N \rightarrow \infty$ will be the same as for $\mathbb{E} \zeta_{1,1}^{(N)} \zeta_{1,2}^{(N)}$ where

$$\zeta_{k,j}^{(N)} := \sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \psi(y) \mathbb{I}\{Y^{j+(k-1)q} = y, |f(X^{j+(k-1)q}) - y| > i\}.$$

Random vectors $(X^1, Y^1), (X^2, Y^2), \dots$ are independent. Therefore, $\mathbb{E} \zeta_{1,1}^{(N)} \zeta_{1,2}^{(N)} = \mathbb{E} \zeta_{1,1}^{(N)} \mathbb{E} \zeta_{1,2}^{(N)}$ and in view of (1) we get

$$\lim_{N \rightarrow \infty} \mathbb{E} \xi_{1,1}^{(N)} \xi_{1,2}^{(N)} = \lim_{N \rightarrow \infty} (\mathbb{E} \zeta_{1,1}^{(N)})^2 = (\text{Err}(f))^2.$$

In a similar way we come to the relation

$$\lim_{N \rightarrow \infty} \mathbb{E} \xi_{1,1}^{(N)} \xi_{2,2}^{(N)} = \lim_{N \rightarrow \infty} (\mathbb{E} \zeta_{1,1}^{(N)})^2 = (\text{Err}(f))^2.$$

Thus $\mathbb{E} X_{N,1} X_{N,2} \rightarrow K (\text{Err}(f))^2$ as $N \rightarrow \infty$. Applying the Lebesgue theorem once again we conclude that

$$\lim_{N \rightarrow \infty} \mathbb{E} (X_{N,j})^2 = \lim_{N \rightarrow \infty} \mathbb{E} (Z_{N,j})^2 = \lim_{N \rightarrow \infty} \left[\mathbb{E} (\zeta_{1,1}^{(N)})^2 + (K-1) \mathbb{E} \zeta_{1,1}^{(N)} \zeta_{1,2}^{(N)} \right],$$

where

$$Z_{N,j} := \frac{1}{\sqrt{K}} \sum_{k=1}^K \zeta_{k,j}^{(N)}, \quad j = 1, \dots, q.$$

Taking into account that $\mathbb{E} \zeta_{1,1}^{(N)} \zeta_{1,2}^{(N)} = (Err(f))^2$ (for each $N \geq 2K$) we get

$$\begin{aligned} \sigma^2 &= \mathbb{E} \left[\sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \psi(y) \mathbb{I}\{Y = y, |f(X) - y| > i\} \right]^2 - (Err(f))^2 \\ &= \mathbb{E} \left[\sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \psi(y) (\mathbb{I}\{Y = y, |f(X) - y| > i\} - \mathbb{P}(Y = y, |f(X) - y| > i)) \right]^2. \end{aligned}$$

To complete the proof we verify condition 3°. Due to the Lebesgue theorem

$$\lim_{N \rightarrow \infty} \text{Cov}(X_{N,1}^2, X_{N,2}^2) = \lim_{k \rightarrow \infty} \text{Cov}(Z_{k,1}^2, Z_{k,2}^2) = 0$$

as $Z_{k,1}$ and $Z_{k,2}$ are independent. Quite similar arguments justify the following relations $\text{Cov}(X_{N,1}X_{N,2}, X_{N,3}X_{N,4}) \rightarrow 0$ and $\text{Cov}(X_{N,1}^2, X_{N,2}X_{N,3}) \rightarrow 0$ as $N \rightarrow \infty$. \square

Let us discuss the established result. Instead of the initial task to study asymptotic behavior of $\sqrt{N}(\hat{T}_N(f) - Err(f))$ we are able to specify the limit law for difference of two estimates of $Err(f)$. Namely, set

$$\hat{L}_{m_N} = \frac{1}{m_N} \sum_{j=1}^{m_N} \frac{1}{\sqrt{K}} \sum_{k=1}^K \sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \left[\hat{\psi}(y, S_k(N)) \cdot \mathbb{I}\{Y^{j+(k-1)q} = y, |f(X^{j+(k-1)q}) - y| > i\} \right]$$

and introduce \hat{L}_q by the same formula with q instead of m_n . Then Lemma 2 affirms that $\sqrt{m_N}(\hat{L}_{m_N} - \hat{L}_q) \xrightarrow{\text{law}} Z_{0,(1-\alpha)\sigma^2} \sim \mathcal{N}(0, (1-\alpha)\sigma^2)$ as $N \rightarrow \infty$. Therefore, if we provide conditions to guarantee that $\sqrt{m_N}(\hat{L}_q - Err(f)) \xrightarrow{\text{P}} 0$ then we can construct the approximate confidence intervals for $Err(f)$. We demonstrate that this is possible for regularized statistics introduced in Bulinski and Rakitko (2014) to identify the significant collections of factors.

For a sequence of random variables $(\eta_N)_{N \in \mathbb{N}}$ we write $\eta_N = O_{\text{P}}(1)$ if, for any $\gamma > 0$, there exists $M(\gamma) > 0$ such that $\mathbb{P}(|\eta_N| \geq M(\gamma)) \leq \gamma$ for all N large enough. Let $(m_N)_{N \in \mathbb{N}}$ be a sequence of positive integers such that $m_N \leq q$ for $q = [N/K]$ and

$$m_N \rightarrow \infty, \quad m_N/N \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

Theorem 2. Let $(m_N)_{N \in \mathbb{N}}$ be a sequence introduced above. Assume that $\varepsilon = (\varepsilon_N)_{N \in \mathbb{N}}$ is a sequence of positive numbers such that $\varepsilon_N \rightarrow 0$ and $m_N^{1/2} \varepsilon_N \rightarrow \infty$ as $N \rightarrow \infty$. Take any vector $\beta = (k_1, \dots, k_r)$ with $1 \leq k_1 < \dots < k_r \leq n$, the corresponding function $f = f^\beta$ and the prediction algorithm defined by $f_{PA} = \hat{f}_{PA, \varepsilon}^\beta$. Let, for any $y \in \mathbb{Y}$ and $k \in \{1, \dots, K\}$, the estimate $\hat{\psi}(y, S_k(N))$ be strongly consistent and

$$\sqrt{\#S_k(N)}(\hat{\psi}(y, S_k(N)) - \psi(y)) = O_{\mathbb{P}}(1), \quad N \rightarrow \infty. \quad (15)$$

Let also (13) hold. Then, as $N \rightarrow \infty$,

$$\sqrt{m_N} \left(\frac{1}{m_N} \sum_{j=1}^{m_N} \frac{1}{\sqrt{K}} \sum_{k=1}^K \sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \left[\hat{\psi}(y, S_k(N)) \mathbb{I}\{A_N(i, j, k, y)\} \right] - \text{Err}(f) \right) \xrightarrow{\text{law}} Z_{0, \sigma^2}.$$

Here $A_N(i, j, k, y) = \{Y^{j+(k-1)q} = y, |f_{PA}(X^{j+(k-1)q}) - y| > i\}$, $Z_{0, \sigma^2} \sim \mathcal{N}(0, \sigma^2)$ and σ^2 was introduced in (14).

Proof. One can show that

$$\begin{aligned} & \sqrt{m_N}(\hat{L}_q - \text{Err}(f)) \\ & - \frac{\sqrt{m_N}}{\sqrt{K}} \sum_{k=1}^K \sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \left[(\hat{\psi}(y, S_k(N)) - \psi(y)) \mathbb{P}(Y = y, |f(X) - y| > i) \right. \\ & \quad \left. + \psi(y) \left(\frac{1}{\#S_k(N)} \sum_{j \in S_k(N)} H_N(y, i, j) \right) \right] \xrightarrow{\mathbb{P}} 0 \end{aligned}$$

as $N \rightarrow \infty$. Here $H_N(y, i, j) = \mathbb{I}\{Y^j = y, |f(X^j) - y| > i\} - \mathbb{P}(Y = y, |f(X) - y| > i)$. For any $i \in \{0, \dots, 2m-1\}$ and $y \in \mathbb{Y}$, the CLT for arrays of i.i.d. random variables with finite second moment implies that

$$\frac{1}{\sqrt{\#S_k(N)}} \sum_{j \in S_k(N)} H_N(y, i, j) = O_{\mathbb{P}}(1), \quad N \rightarrow \infty.$$

Since $m_N/\#S_k(N) \rightarrow 0$ we get

$$\frac{\sqrt{m_N}}{\sqrt{K}} \sum_{k=1}^K \sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \psi(y) \frac{1}{\#S_k(N)} \sum_{j \in S_k(N)} H_N(y, i, j) \xrightarrow{\mathbb{P}} 0, \quad N \rightarrow \infty.$$

In a similar way in view of (15) one has

$$\frac{\sqrt{m_N}}{\sqrt{K}} \sum_{k=1}^K \sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} (\hat{\psi}(y, S_k(N)) - \psi(y)) \mathbf{P}(Y = y, |f(X) - y| > i) \xrightarrow{\mathbf{P}} 0, \quad N \rightarrow \infty.$$

Thus under conditions of Theorem 2 the asymptotic behavior of $\sqrt{m_N}(\hat{L}_{m_N} - \hat{L}_q)$ is the same as for $\sqrt{m_N}(\hat{L}_{m_N} - \text{Err}(f))$. \square

In Velez et al. (2007) the following choice of the penalty function ψ was proposed

$$\psi(y) = \frac{c}{\mathbf{P}(Y = y)}, \quad y \in \mathbb{Y}, \quad c = \text{const} > 0.$$

This choice was justified in Bulinski (2012) for binary response Y . We will employ this penalty function for nonbinary response as well, i.e. when $\mathbb{Y} = \{-m, \dots, 0, \dots, m\}$. Further we assume that $\mathbf{P}(Y = y) > 0$ for all $y \in \mathbb{Y}$ and w.l.g. $c = 1$.

Introduce $A_N(y, S_k(N)) = \{Y^j \neq y, j \in S_k(N)\}$, $N \in \mathbb{N}$, $k \in \{1, \dots, K\}$, $y \in \mathbb{Y}$ and set (as usual $0/0 := 0$)

$$\hat{\mathbf{P}}_{S_k(N)}(Y = y) := \frac{\sum_{j \in S_k(N)} \mathbb{I}\{Y^j = y\}}{\sharp S_k(N)}, \quad \hat{\psi}(y, S_k(N)) := \frac{\mathbb{I}\{\Omega \setminus A_N(y, S_k(N))\}}{\hat{\mathbf{P}}_{S_k(N)}(Y = y)}. \quad (16)$$

Corollary 1. *The estimate $\hat{\psi}$ defined by (16) satisfies conditions of Theorem 2.*

Proof. Fix arbitrary $y \in \mathbb{Y}$ and $k = 1, \dots, K$. One can easily check that $\hat{\psi}(y, S_k(N))$ is a strongly consistent estimate of $\psi(y)$. Moreover, by CLT for arrays of i.i.d. random variables we have

$$\sqrt{\sharp S_k(N)}(\hat{\mathbf{P}}_{S_k(N)}(Y = y) - \mathbf{P}(Y = y)) \xrightarrow{\text{law}} Z_{0, \sigma_1^2(y)} \sim \mathcal{N}(0, \sigma_1^2(y)), \quad N \rightarrow \infty,$$

where $\sigma_1^2(y) = \mathbf{P}(Y = y)(1 - \mathbf{P}(Y = y))$. Taking into account that $\hat{\mathbf{P}}_{S_k(N)}(Y = y) \rightarrow \mathbf{P}(Y = y)$ a.s. and $\sqrt{\sharp S_k(N)}\mathbb{I}\{A_N(y, S_k(N))\} \xrightarrow{\mathbf{P}} 0$ as $N \rightarrow \infty$, one can write by Slutsky's lemma that

$$\sqrt{\sharp S_k(N)}(\hat{\psi}_{S_k(N)}(y) - \psi(y)) \xrightarrow{\text{law}} Z_{0, \sigma_2^2(y)} \sim \mathcal{N}(0, \sigma_2^2(y)), \quad N \rightarrow \infty,$$

where $\sigma_2^2(y) = (1 - \mathbf{P}(Y = y))\mathbf{P}(Y = y)^{-3}$. Thus (15) holds. Now we verify (13). Clearly, $\hat{\psi}(y, S_k(N)) \leq \sharp S_K(N)$ for any $N \in \mathbb{N}$. Put $\varepsilon := \min_{y \in \mathbb{Y}} \mathbf{P}(Y = y)$. Then by the Hoeffding inequality

$$\mathbf{E}|\hat{\psi}(y, S_k(N))|^4 = \mathbf{E}\left[|\hat{\psi}_{N,k}(y)|^4 \mathbb{I}\left\{|\hat{\mathbf{P}}_{S_k(N)}(Y = y) - \mathbf{P}(Y = y)| > \varepsilon/2\right\}\right]$$

$$\begin{aligned}
& +\mathbb{E}\left[(\widehat{\psi}(y, S_k(N)))^4 \mathbb{I}\left\{\left|\widehat{\mathbb{P}}_{S_k(N)}(Y=y) - \mathbb{P}(Y=y)\right| \leq \varepsilon/2\right\}\right] \\
& \leq 2(\#S_K(N))^4 \exp\{-\#S_1(N)\varepsilon^2/2\} + 2^4/\varepsilon^4.
\end{aligned}$$

Thus we come to (13). \square

To simplify notation we will write in the following theorem $\widehat{Err}_K(f_{PA}, \xi_N)$ for random variable introduced in (2) replacing $\widehat{\psi}(y, S_k(N))$ by $\widehat{\psi}(y, \overline{S_k(N)})$, $y \in \mathbb{Y}$, $k = 1, \dots, K$. After such replacement in (4) – (6) we obtain the new row-wise exchangeable array $\{X_{N,j}, 1 \leq j \leq q, N \in \mathbb{N}\}$ and therefore all established results hold true in this case.

Theorem 3. *Let $\varepsilon_N \rightarrow 0$ and $N^{1/2}\varepsilon_N \rightarrow \infty$ as $N \rightarrow \infty$. Then, for each $K \in \mathbb{N}$, any vector $\beta = (k_1, \dots, k_r)$ with $1 \leq k_1 < \dots < k_r \leq n$, the corresponding function $f = f^\beta$ and prediction algorithm defined by $f_{PA} = \widehat{f}_{PA, \varepsilon}^\beta$, the following relation holds:*

$$\sqrt{N}(\widehat{Err}_K(f_{PA}, \xi_N) - Err(f)) \xrightarrow{law} Z_{0, \sigma^2} \sim \mathcal{N}(0, \sigma^2), \quad N \rightarrow \infty. \quad (17)$$

Here σ^2 is variance of the random variable

$$V = \sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \frac{\mathbb{I}\{Y=y\}}{\mathbb{P}(Y=y)} (\mathbb{I}\{|f(X) - y| > i\} - \mathbb{P}(|f(X) - y| > i | Y=y)). \quad (18)$$

Proof. Set, for $f : \mathbb{X} \rightarrow \mathbb{Y}$ and $N \in \mathbb{N}$,

$$T_N(f) := \frac{1}{K} \sum_{k=1}^K \frac{1}{\#S_k(N)} \sum_{i=1}^{2m-1} \sum_{i-m < |y| \leq m} \psi(y) \sum_{j \in S_k(N)} \mathbb{I}\{Y^j = y, |f(X^j) - y| > i\}.$$

The Slutsky lemma shows that the limit behavior of the random variables introduced in (3) will be the same as for random variables

$$\begin{aligned}
\rho_N &:= \sqrt{N}(T_N(f) - Err(f)) \\
&= -\frac{\sqrt{N}}{K} \sum_{k=1}^K \sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \frac{(\widehat{\mathbb{P}}_{S_k(N)}(Y=y) - \mathbb{P}(Y=y))\mathbb{P}(Y=y, |f(X) - y| > i)}{\mathbb{P}(Y=y)^2} \\
&= \frac{\sqrt{N}}{K} \sum_{k=1}^K \sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \frac{1}{\#S_k(N)} \sum_{j \in S_k(N)} \frac{\mathbb{I}\{Y^j = y, |f(X^j) - y| > i\}}{\mathbb{P}(Y=y)}
\end{aligned}$$

$$-\frac{\sqrt{N}}{K} \sum_{k=1}^K \sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \frac{1}{\# \overline{S_k(N)}} \sum_{j \in S_k(N)} \mathbb{I}\{Y^j = y\} \frac{\mathbb{P}(Y = y, |f(X) - y| > i)}{(\mathbb{P}(Y = y))^2}.$$

Let $a_k, b_k, k = 1, \dots, K$, be any real numbers. We use the following simple observation

$$\frac{1}{\# S_k(N)} \sum_{k=1}^K a_k + \frac{1}{\# \overline{S_k(N)}} \sum_{l=1, \dots, K; l \neq k} b_l = \sum_{k=1}^K \left(\frac{a_k}{\# S_k(N)} + b_k \sum_{l=1, \dots, K; l \neq k} \frac{1}{\# \overline{S_l(N)}} \right).$$

Combining the latter formulas we can write

$$\rho_N = \frac{\sqrt{N}}{K} \sum_{k=1}^K \sum_{j \in S_k(N)} \left(\frac{V_1^j}{\# S_k(N)} + V_2^j \sum_{l=1, \dots, K; l \neq k} \frac{1}{\# \overline{S_l(N)}} \right)$$

where

$$V_1^j = \sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \frac{\mathbb{I}\{Y^j = y, |f(X^j) - y| > i\}}{\mathbb{P}(Y = y)},$$

$$V_2^j = - \sum_{i=0}^{2m-1} \sum_{i-m < |y| \leq m} \frac{\mathbb{I}\{Y^j = y\} \mathbb{P}(Y = y, |f(X) - y| > i)}{(\mathbb{P}(Y = y))^2}.$$

Take any $k \in \{1, \dots, K\}$ and employ CLT for an array of bounded centered i.i.d. random variables $\{V_1^j + V_2^j, j \in S_k(N), N \in \mathbb{N}\}$. Then

$$\frac{1}{\sqrt{\# \overline{S_k(N)}}} \sum_{j \in S_k(N)} (V_1^j + V_2^j) \xrightarrow{law} Z_{0, \sigma^2} \sim \mathcal{N}(0, \sigma^2), \quad N \rightarrow \infty,$$

where $\sigma^2 = \text{var}(V_1^j + V_2^j)$. Note now that, for each $k = 1, \dots, K$,

$$\frac{N}{\# S_k(N)} \rightarrow \frac{1}{K}, \quad \sum_{l=1, \dots, K; l \neq k} \frac{N}{\# \overline{S_l(N)}} \rightarrow \frac{1}{K}, \quad N \rightarrow \infty.$$

For each $N \in \mathbb{N}$, the families of random variables $\{V_1^j + V_2^j, j \in S_k(N)\}, k = 1, \dots, K$, are independent. Thus we come to the following relation

$$\rho_N \xrightarrow{law} Z_{0, \sigma^2} \sim \mathcal{N}(0, \sigma^2), \quad N \rightarrow \infty.$$

Obviously we can write $\sigma^2 = \text{var } V$ where V is introduced in (18). The proof is complete. \square

Remark 3. It is not difficult to construct the consistent estimates $\hat{\sigma}_N$ of unknown σ appearing in (17). Therefore (if $\sigma^2 \neq 0$) we can claim that under conditions of Theorem 3

$$\frac{\sqrt{N}}{\hat{\sigma}_N} (\hat{Err}_K(f_{PA}, \xi_N) - Err(f)) \xrightarrow{law} \frac{Z}{\sigma} \sim \mathcal{N}(0, 1), \quad N \rightarrow \infty.$$

Remark 4. It is interesting to compare Theorem 3 with simulations corresponding to Example 3 when N is rather small (e.g., $N = 200$). In this case the choice of $\widehat{\psi}(y, \overline{S_k(N)})$ can lead to better identification of significant factors.

BIBLIOGRAPHY

Arlot, S., Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statist. Surv.* **4**, 40–79.

Berti, Patrizia; Pratelli, Luca; Rigo, Pietro. (2004). Limit theorems for a class of identically distributed random variables. *Ann. Probab.* **32**, no. 3, 2029–2052.

Billingsley P. (1968). *Convergence of Probability Measures*. John Wiley & Sons Inc., New York.

Bulinski, A.V. (2014). On foundation of the dimensionality reduction method for explanatory variables. *J. Math. Sci.*, DOI 10.1007/s10958-014-1838-7.

Bulinski, A.V. (to appear, 2014). Central limit theorem related to MDR method.: Proceedings of the Fields Institute International Symposium on Asymptotic Methods in Stochastics, in Honour of Miklós Csörgő's Work on the occasion of his anniversary. arXiv:1301.6609 [math.PR].

Bulinski, A., Butkovsky, O., Sadovnichy, V., Shashkin, A., Yaskov, P., Balatskiy, A., Samokhodskaya, L., Tkachuk, V. (2012). Statistical methods of SNP data analysis and applications. *Open Journal of Statistics*. **2**(1), 73–87.

Bulinski A.V., Rakitko A.S. (2014). Estimation of nonbinary random response. *Dokl. Math.*, **455**(6), 1–5.

Chernoff, H., Teicher, H. (1958). A Central Limit Theorem for Sums of Interchangeable Random Variables. *Ann. Math. Statist.* **29**(1), 118–130.

- Golland, P., Liang, F., Mukherjee, S., Panchenko, D. (2005). *Permutation Tests for Classification*. LNCS, **3559**, 501–515.
- Hastie T., Tibshirani R. and Friedman J. (2008). *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. Springer, New York. Second edition.
- Lee S., Epstein M.P., Duncan R. and Lin X. (2012). Sparse Principal Component Analysis for Identifying Ancestry-Informative Markers in Genome Wide Association Studies. *Genet. Epidemiol.* **36**, 293–302.
- Moore J.B., , Asselbergs F.W. and Williams S.M. (2010). Bioinformatics challenges for genome-wide association studies. *Bioinformatics.* **26**, 445–455.
- Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, R.L., Dupont, W.D., Parl, F.F., Moore, J.H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum Genet.* **69**(1), 138–147.
- Röllin, A. (2013). Stein’s method in high dimensions with applications. *Ann. Inst. Henri Poincaré Probab. Stat.* **49**(2), 529–549.
- Schwender, H., Ruczinski, I. (2010). Logic regression and its extensions. *Adv. Genet..* **72**, 25–45.
- Sikorska K., Lesaffre E., Groenen P.F.G., and Eilers P.H.C. (2013). GWAS on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies. *BMC Bioinformatics*, **14**:166.
- Tibshirani R.J. and Taylor J. (2012). Degrees of freedom in lasso problems. *Ann. Statist.* **40**, 1198–1232.
- Velez, D.R., White, B.C., Motsinger, A.A., Bush, W.S., Ritchie, M.D., Williams, S.M., Moore, J.H. (2007). A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet. Epidemiol..* **31**(4), 306–315.

Visscher, P.M., Brown, M.A., McCarthy, M.I., Yang, J. (2012). Five Years of GWAS Discovery. *Am. J. Hum. Genet.* **90**, 7–24.

Weber N.C. (1980). A martingale approach to central limit theorems for exchangeable random variables. *J. Appl. Probab.*, **17**(3), 662-673.